USING THE KTH NEAREST NEIGHBOR CLUSTERING PROCEDURE TO DETERMINE THE NUMBER OF SUBPOPULATIONS

BY

M. Anthony Wong and Christian Schaak

Working Paper #1338-82

USING THE KTH NEAREST NEIGHBOR CLUSTERING PROCEDURE TO DETERMINE THE NUMBER OF SUBPOPULATIONS

BY

M. Anthony Wong and Christian Schaak

Working Paper #1338-82

AUTHORS' FOOTNOTE

M. Anthony Wong is Assistant Professor, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139. Christian Schaack is a graduate assistant in the Operations Research Center at M.I.T. This research was supported in part by the Sloan School of Management Faculty Research Funds.

ABSTRACT

A major problem in cluster analysis is determining the number of subpopulations from the sample data. In this study, it is assumed that the subpopulations correspond to modes of the population density function. The kth nearest neighbor clustering procedure, which is known to be set-consistent for high-density clusters, is then shown to be useful in providing: (1) a diagnostic plot which will indicate the number of subpopulations present, and (2) a bootstrap procedure for testing the existence of two or more subpopulations. The performance of these procedures will be illustrated by real examples.

KEYWORDS: Modes; kth nearest neighbor clustering; diagnostic plot; hypothesis testing; bootstrap.

1. INTRODUCTION

1.1 Background

A recent study by Blashfield and Aldenderfer (1978) shows that numerous clustering methods have been developed in the past two decades. A review of many of these techniques can be found in Cormack (1971), Anderberg (1973), Sneath and Sokal (1973), Everitt (1974), Hartigan (1975), and Spath (1980). The validity of the sample clusters obtained by these methods is always questionable, however, due to the lack of development in the probabilistic and statistical aspects of clustering methodology. Consequently, the existing clustering procedures are often regarded as heuristics generating artificial clusters from a given set of sample data, and there is a need of clustering procedures that are useful for drawing statistical inferences about the underlying population from a sample. In this paper, we consider the important problem of assessing and testing the number of "clusters" or "subpopulations" present in the population.

1.2 Statistical Inference Under the Density-contour Clustering Model

In this study, we assume that the clustering data consist of a sample from a distribution F with density function f, on which population clusters are defined by a clustering model. The clustering model that will be used here is the "density-contour" model given in Hartigan (1975) and Wong and Lane (1981). Using this model, the true population clusters can be defined on f as follows: for all $f^* > 0$, a density-contour cluster at level f^* in the population is defined as a maximal connected set of the form $\{x \mid f(x) \geq f^*\}$. The family T of such clusters forms a tree in the sense that $A \varepsilon T$, $B \varepsilon T$ implies either $A \supset B$, $B \supset A$, or $A \cap B = \phi$, the empty set.

A hierarchical clustering procedure, which produces a sample clustering tree T_N on the observations $X_1, \ldots, X_N$ may then be evaluated by examining whether T_N converges to T with probability one when N approaches infinity. A clustering method (or equivalently, T_N) is said to be strongly set-consistent for density-contour clusters (or T) if for any A, $B \varepsilon T$, $A \cap B = \phi$,

$$\Pr \{ A_N \cap B_N = \phi \text{ as } N \to \infty \} = 1,$$

where A_N and B_N are respectively the smallest cluster in the sample tree T_N containing all the sample points in A and B. Since $A \supset B$ implies $A_N \supset B_N$, this limit result means that the tree relationship in T_N converges strongly to the tree relationship in T. This consistent cluster-estimation problem under the density-contour clustering model has been addressed by Hartigan (1981) and Wong and Lane (1981). Hartigan (1981) has shown that most of the best known hierarchical clustering methods are not set-consistent, while Wong and Lane (1981) developed a kth nearest neighbor clustering procedure which is strongly set-consistent for density-contour clusters.

The problem of hypothesis testing under the density-contour clustering model did not receive much attention. (See however, Hartigan (1977) for a discussion of the DIP statistic for testing bimodality.) One important feature of the density function f under the density-contour clustering model is the modes of f, each of which is the limit of a decreasing sequence of density-contour clusters. In this paper, it is assumed that any subpopulation in the population corresponds to a mode in the density function f. Our aim is to develop procedures that are useful for assessing and testing the number of modes present in f. It will be shown that the kth nearest neighbor clustering procedure given in Wong and Lane (1981) is useful in providing (i) a diagnostic plot for assessing the number of modes, and (ii) a statistic for testing multimodality.

Using the above formulation, the statistical problem being considered is that of determining the number of modes in the underlying density f. A brief review of the literature on testing for modes will be given in Section 2. In Section 3, the kth nearest neighbor clustering procedure will first be reviewed, and then it will be shown how a diagnostic plot based on this procedure can be constructed to assess multimodality. A test statistic for examining multimodality is proposed in Section 4, and it will also be shown how the significance level of a sample test statistic can be estimated by using the bootstrap procedure. Generated data will be used to illustrate the performances of these procedures. And in Section 5, the practical utility of the proposed procedures are demonstrated by several well-known data sets.

2. LITERATURE REVIEW

Several authors have studied the problem of testing for clusters. In Engelman and Hartigan (1969) and Hartigan (1978), a likelihood ratio approach to the problem of testing whether the data indicate the presence of two different univariate normal populations or only one is proposed. Multivariate generalizations of Engelman and Hartigan's work can be found in Lee (1979), in which the union-intersection principle of test construction is used to develop a multivariate test for clusters. One major drawback of these testing procedures is that they are based on a restrictive parametric clustering model, where clusters are assumed to be components of a normal mixture.

In this paper, the nonparametric density-contour clustering model is used, where clusters are defined by the density contours of the underlying density function. Our aim is to develop procedures that are useful for assessing and testing multimodality. In the clustering literature, two different statistics have been proposed for testing bimodality in one dimension. Kruskal's test given in Giacomelli et. al. (1971) is based on the differences between order statistics, while Hartigan's (1977) DIP statistic looks for a large interval between two sets of small intervals in the minimum spanning tree obtained for the sample observations. The major problem encountered in using these test statistics is the selection of an appropriate distribution function for the null hypothesis. The uniform and normal distributions have been used by both of the above authors to compute the sampling distribution of the proposed statistics, but the appropriateness of using these null distributions in cluster anlysis remains questionable.

Silverman (1981) proposed a statistic for testing the multimodality of an underlying density function f which is based on the kernel density estimate. More importantly, he proposed an intuitively appealing bootstrap procedure for estimating the significance level of a sample value of his statistic, without having to use the uniform or normal as the null distribution. In this paper, a statistic is proposed for testing multimodality, which is based on the kth nearest neighbor clustering procedure given in Wong and Lane (1981), and it will be shown how a modified bootstrap procedure can be used to estimate the significance level of a sample value of this statistic.

3. A DIAGNOSTIC PLOT FOR THE NUMBER OF MODES

3.1 The kth Nearest Neighbor Clustering Procedure

In this section, it will be shown that the kth nearest neighbor clustering procedure given in Wong and Lane (1981) can be used to provide a diagnostic plot for assessing the number of modes in a density f using some sample data $X_1, X_2, \ldots, X_N$ from f. This clustering procedure can be described as follows:

Step 1: For $i = 1, 2, \ldots, N$, compute $d_k(X_i)$, the kth nearest neighbor distance for observation X_i.

Step 2: Compute the distance matrix D as follows:

$$
\begin{aligned}
D(X_i, X_j) &= 0, \quad \text{if } X_i = X_j; \\
&= 1/2\,[d_k(X_i) + d_k(X_j)], \quad \text{if } d^*(X_i, X_j) \le d_k(X_i) \\
&\qquad\qquad \text{or } d^*(X_i, X_j) \le d_k(X_j), \\
&\qquad\qquad \text{where } d^* \text{ is the Euclidean metric:} \\
&= \infty, \quad \text{otherwise.}
\end{aligned}
$$

Step 3: Apply the single linkage clustering algorithm to the computed distance matrix D to obtain the sample tree of high-density clusters.

3.2 A Diagnostic Plot for the Number of Modes

In Wong and Lane (1981), it is pointed out that for the kth nearest neighbor clustering procedure to be stongly set-consistent, k has to be chosen in such a way that $k(N)/N \to 0$, and $k(N)/\log N \to \infty$, as $N \to \infty$. However, the problem of choosing k in practice has not been dealt with, although it has been suggested that a range of values of k should be tried. Here, it is proposed that the number of modes identified in the sample hierarchical clustering when different values of k are used should be plotted against k because this plot is useful in suggesting the number of modes in the population.

It is not difficult to see that the value of k controls the amount by which the data are smoothed to give the density estimate on which the clustering procedure is based. When k increases from 1 to N, the density estimate becomes smoother or less bumpy; that is, the number of identified modes is a non-increasing function of k. (This result is proved in Silverman (1981) for the kernel density estimate.) Hence, the plot of "number of estimated modes" against k will show a non-increasing step function; and it is expected that when the number of estimated modes reaches the true number of modes, it will be stable over a range of values of k. The results of a Monte Carlo study performed to examine the effectiveness of this diagnostic plot will be reported next.

3.3 Empirical Illustrations of the Diagnostic Plot

Sixteen experiments were run using data generated from various normal distributions and mixtures thereof. The four diagnostic plots shown in

Figure A are obtained for two samples of size 50 and two of size 100 that are generated according to the univariate unimodal standard normal distribution, N(0,1), while those shown in Figure B are obtained for corresponding samples generated according to the bivariate unimodal normal distribution, BVN $[(0,0), \binom{10}{01}]$. In all of these plots, a very extensive plateau can be observed where the number of identified modes is 1, while no other stable number of modes is indicated.

Figures C and D show some interesting, yet disturbing features of the proposed diagnostic plot. Although, as can be expected of samples from bimodal distributions, all of the plots show a wide range of stability for bimodality, some of the plots also show stable plateaus for trimodality (see Figures C(a2), C(b2), and D(b2)). Since each of the samples used to obtain Figure C(b1) and C(b2) consists of 30 observations from N(0,1) and 70 observations from N(8,4), it is unreasonable to expect the number of identified modes to be greater than 1 when k is greater than 30. Hence, the relatively short bimodality plateau shown in Figure C(b2) is not unexpected. However, the diagnostic plots shown in Figures C(b1) and C(b2) also show that two different samples from the same distributions can give plateaus of fairly different widths.

It is difficult to account for the trimodality plateau that is evident in Figure C(a2), but at least in this case it is significantly narrower than the very stable bimodality plateau. For Figure D(b2), a look at the corresponding scatterplot (Figure E) suggests that the appearance of a sizeable trimodality plateau in the diagnostic plot is not unreasonable.

In this section, we have shown that the proposed diagnostic plot is useful in indicating the number of modes that are present in a population.

It is also useful in suggesting the possible existence of finer subpopulations. It is however, sensitive to the sample sizes from different subpopulations, but only in as much as they impose upper bounds on the width of the plateaus. On the whole, the proposed plot seems to be a valuable diagnostic tool for assessing multimodality.

4. A TEST STATISTIC FOR TESTING THE MULTIMODALITY OF A UNIVARIATE DENSITY f

4.1 The Test Statistic

Investigation of the number of modes or maxima in a density has been considered by several authors, for example Good and Gaskins (1980) and Silverman (1981). As remarked by Silverman (1981), it is unfortunate that most of the proposed methods seem to depend on some arbitrary implicit or explicit choice of the scale of the effects being studied. The simple approach based on the kth nearest neighbor clustering procedure described in this paper has the virtue of making this choice in an automatic and natural way.

A possible test statistic for hypotheses concerning the number of modes in a univariate density f can be obtained by applying the kth nearest neighbor clustering procedure to the sample data from f. Now, the value of k controls the amount by which the data are smoothed to obtain the density estimate on which the clustering procedure is based. Therefore, for example, if the data are strongly bimodal, a large value of k will be needed to give a sample hierarchical clustering with only one mode. Suppose that we wish to test the null hypothesis that the density f underlying the data has M modes, against the alternative that f has more than M modes. Then define the critical k-value, k_{crit}, by

$k_{crit} = \inf \{k; \hat{f}(\cdot, k)$ has at most M modes$\}$ where $\hat{f}(\cdot, k)$ is the density estimate obtained by the kth nearest neighbor procedure. Large values of k_{crit} will reject the null hypothesis.

4.2 Assessing the Significance Level P of a Sample Value of k_{crit}

Suppose that the value of k_{crit} obtained from the sample data is k_o for testing $H_o : f$ has M modes against $H_A : f$ has more than M modes. Our aim is to estimate the observed significance level

$$P = P_r \{k_{crit} > k_o \mid H_o \text{ is true}\},$$

so that we can reject H_o when P is sufficiently small. It is shown below how an estimate of P can be obtained by using a bootstrap procedure (See Efron, 1979).

To obtain a conservative estimate of P, an appealing choice for the null distribution f_o, from which simulated samples are to be taken, is the density estimate obtained when k_o is used as the value of the parameter k, scaled to have variance equal to the sample variance S^2 of the data. For univariate data, it is easy to simulate from f_o by using the bootstrap method. As pointed out in Efron (1979), N independent observations from f_o are given by

$$y_i = \left(1 + \frac{d_{k_o}^2(X_{I(i)})}{3s^2}\right)^{-1/2} (X_{I(i)} + d_{k_o}(X_{I(i)})\mu_i[-1,1]),$$

where $X_{I(i)}$ are sampled uniformly, with replacement, from the data $X_1, X_2, \ldots, X_N$; s^2 is the sample variance of the data, $d_{k_o}(X_{I(i)})$ is the k_oth nearest neighbor distance of observation $X_{I(i)}$, and $\mu_i[-1,1]$ is an independent sequence of uniform random variables distributed between -1 and +1. And the value of P can then be estimated by finding the proportion of R bootstrap samples of size N which give values of k_{crit} greater than k_o.

The computational procedure can be summarized as follows:

Step 0: Compute s^2 and find k_o for the sample data.

Step 1: For $i = 1$ to N,

sample with replacement from $\{1, 2, \ldots, N\}$;

let I(i) be the ith pick, and let

$$y_i = \left(\frac{d_{k_o}^2 (X_{I(i)})}{3s^2} \right)^{-1/2} (X_{I(i)} + d_{k_o}(X_{I(i)}) \mu_i[-1,1]).$$

Step 2: Apply the kth nearest neighbor clustering procedure, with $k = k_o$, to the bootstrapped data $y_1, y_2, \ldots, y_N$. Test if the number of sample-modes SM is greater than M.

Step 3: Repeat steps 1 and 2 R times (we will use R=120).

Step 4: Let the estimate of $P = \frac{\#\text{ times that } (SM > M)}{120}$.

Then, H_o is accepted at the 5% level if the p-value P is greater than 0.05.

It should be borne in mind that this test is very conservative as it uses the most extreme k that yields M-modality for the sample

$X_1, X_2, \ldots, X_N$.

We applied the above test to various univariate normal distributions and mixtures thereof. Twenty-five samples of size 100 were taken for each distribution studied, the results of the test for various null hypotheses (one, two, and three modes) are given in Table 1 below; they consist of values of k_o (the value of k_{crit} obtained from the sample) and the corresponding estimates of P. Note that these results must of course be interpreted as a hierarchical set of significance tests: If (M-1) - modality is not rejected by the test, then there is no point in testing for M-modality. So we should test successively for an increasing number of modes until we find a number that is accepted. In the following discussion, we will use a significance level of 5%.

Table 1(a) shows that none of the 25 samples fron N(0,1) leads to a rejection of a unimodal null hypothesis. Equally encouraging are the results for the fifty-fifty mixture of N(0,1) and N(4,1); bimodality is rejected only once out of twenty-five samples. Moreover, the empirical power of testing "H_o: the distribution is unimodal", against "H_a: the distribution is bimodal" for samples from this mixture is very good: 92%

For the trimodal mixture in Table 1(b) (25 observations from N(0,1), 25 from N(4,1) and 50 from N(8,1)), the test fails to reject unimodality in 21 cases out of 25; and in two out of the remaining four cases, bimodality cannot be rejected. The reason for the poor performance of the proposed test for this mixture is thought to lie primarily in the small (25 observations) and uneven (25/25/50) subsample sizes. Indeed the density estimate turns unimodal for k_o around 25 because of the small subsample sizes, but for $k_o = 25$ (small with respect to the sample size of 100), the density estimate is still very sensitive to perturbations around the sample points, and as perturbations are exactly what the bootstrap does, the bootstrapped sample is most likely to be multimodal for k_o; hence the test tends to accept unimodality. It should again be pointed out that the proposed test is hierarchical in nature, and there is little point in testing for bimodality if unimodality cannot be rejected.

For the results in Table 1(c),(25 observations from N(0,1), 50 for N(4,1), 25 from N(8,1)), unimodality cannot be rejected for any of the 25 samples, so indeed we have a very conservative test.

[Table 1 about here]

The proposed test statistic has been shown to perform well in one dimension for truly unimodal distributions and for bimodal distributions with nicely separated modes of equal importance. It behaves comparatively poorly when the subsample sizes are small and/or uneven. In fact, it is more conservative than expected, and needs to be improved if it is to be a sharp testing tool; especially since its computational expenditure is non-negligible (on the average, about one hour of CPU-time is consumed on a Prime 850, for a program that tests for 1,2,3, and 4 modes using a sample of size 100, i.e., roughly a quarter of an hour of CPU-time per null hypothesis tested). Moreover, although the proposed test statistic is also well-defined for multivariate data, the bootstrap procedure described above for estimating the p-value of a sample test statistic cannot be easily generalized to several dimensions. Hence, much work has yet to be done to develop an appropriate generalization of this testing procedure.

5. ILLUSTRATIVE EXAMPLES

In this section, the effectiveness of the proposed diagnostic plot and the testing procedure are illustrated with real examples.

The real univariate data sets used are:

(1) the chondrite data from Good and Gaskins (1981), 22 observations.

(2) the petal lengths of Fisher's Iris data, (Fisher, 1936), for two Iris species (setosa and versicolor), 100 observations, (2x50)

(3) the petal lengths of Fisher's Iris data for three Iris species (Setosa, versicolor and virginica), 150 observations (3x50).

- Data Set (1)

We have analyzed the data which consist of the distribution of silica in 22 chrondrite meteors; this data has been studied previously, among others, by Good and Gaskins (1981), and Silverman (1981).

Table 2

Percentages y Silica in 22 Chondrites

y	20.77	22.56	22.71	22.99	26.39	27.08
y	27.32	27.33	22.57	27.81	28.69	29.36
y	30.25	31.89	32.88	33.23	33.28	33.40
y	33.52	33.83	33.95	34.82		

The diagnostic plot (Figure F(a)) reveals only one very stable plateau for unimodality. Small plateaus are also detected for two and three modes.

The testing procedure developed in Section 4 yields the following k_{crit} statistics:

H_o	k_o	Estimated $P_r(k_{crit} > k_o)$
unimodal	8	0.067
bimodal	5	0.677
trimodal	2	0.833

Consequently, we cannot reject unimodality at the 5% level. (Note that we can at the 10% level, in which case we accept bimodality of the population). We cannot accept trimodality exclusive of uni- or bimodality, which is not surprising considering the small number of observations; they could be sampled from any distribution. We do find a finer trimodal substructure, indicated by the diagnostic plot, but no more than an indication of it (see also Good and Gaskins (1981), and Silverman (1981) whose conclusion is questionable).

-Date set (2) consists of the petal lengths of two iris species, setosa and versicolor.

The diagnostic plot (Figure F(b)) reveals a stable plateau at 2 modes, but also suggests small three- and four-mode plateaus. It seems to indicate a basic bimodal population with possibly some finer substructures (small additional modal regions).

When we tested for various numbers of modes, we obtained the following results:

H_o	k_o	estimated $P_r(k_{crit} > k_o)$
unimodal	50	0.000
bimodal	19	0.025
trimodal	13	0.017
quadimodal	7	0.583

Hence we reject the first three null hypotheses at the 5% level, and accept quadrimodality. It is known that the petal lengths between the two species are very different, but the test seems to indicate that the distribution might have four modes.

-Data set (3) includes three species of iris (setosa, versicolor and virginica).

The diagnostic plot (Figure F(c)) shows a stable two-mode plateau, and also a relatively stable plateau at five modes; however, the plateaus found for two and three modes shown in Figure F(b) have virtually disappeared.

The test statistic also yields an altogether different picture.

H_o	k_o	estimated P_r $(k_{crit} > k_o)$
unimodal	51	0.750
bimodal	19	0.325
trimodal	16	0.108
q-modal	14	0.008

The test does not reject unimodality. Now, it is known that the Iris Setosa species is very different from the other two which are not distinct from one another; so why does the test not reject unimodality? The main culprit seems to be the fact that the two modes one expects to find are of uneven sizes, and the test is very sensitive to uneven subsample sizes as seen earlier on.

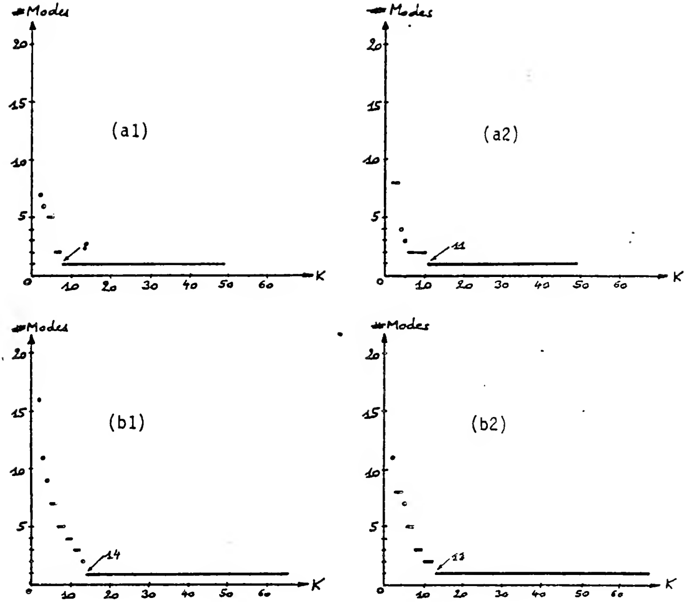

FIGURE A

Diagnostic Plots: Normal Distributions

(a1) and (a2): 50 observations from N(0,1)
(b1) and (b2): 100 observations from N(0,1)

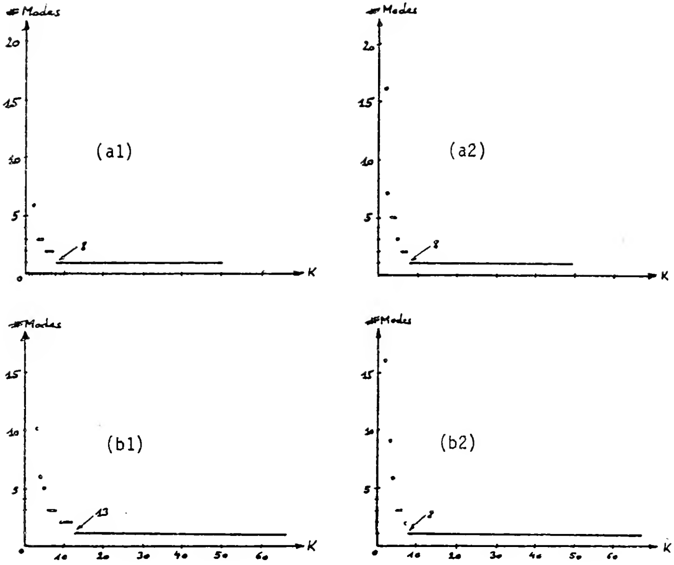

FIGURE B

Diagnostic Plots: Normal Distributions

(a1) and (a2): 50 observations from $BVN[(0,0)\binom{10}{01}]$

(b1) and (b2): 100 observations from $BVN[(0,0)\binom{10}{01}]$

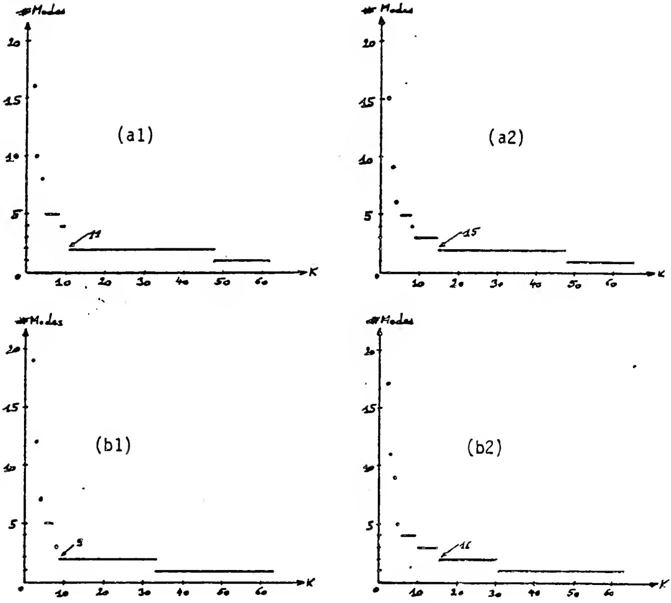

FIGURE C

Diagnostic Plots: Normal Mixtures

(a1) and (a2): 50 observations from N(0,1) and 50 from N(5,1)
(b1) and (b2): 30 observations from N(0,1) and 70 from N(8,4)

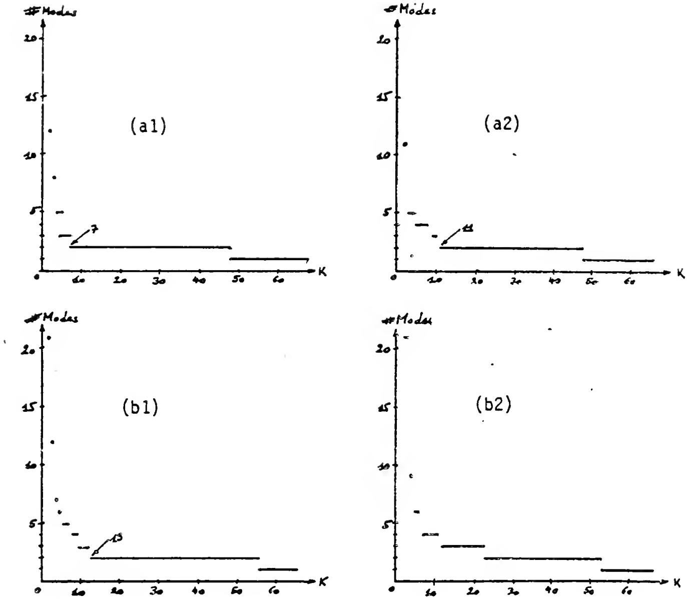

FIGURE D

Diagnostic Plots: Normal Mixtures

(a1) and (a2): 50 observations from BVN$[(0,0),\binom{10}{01}]$ and 50 from BVN $[(5,0),\binom{10}{01}]$

(b1) and (b2): 50 observations from BVN$[(0,0),\binom{10}{01}]$ and 100 from BVN$[10,0),\binom{90}{01}]$

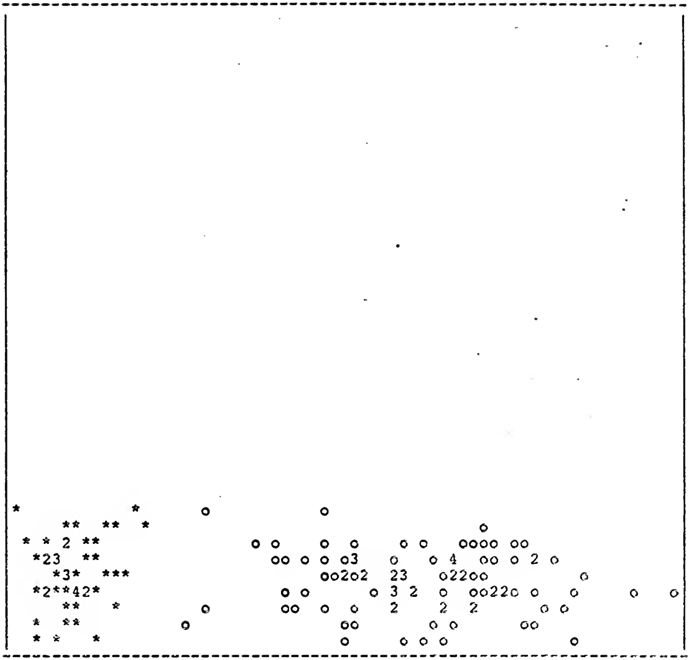

FIGURE E

xy-Plots: Normal Mixtures

50 observations from BVN[(0,0),$\begin{pmatrix}10\\01\end{pmatrix}$ and 100 from BVN[(10,0),$\begin{pmatrix}90\\01\end{pmatrix}$]

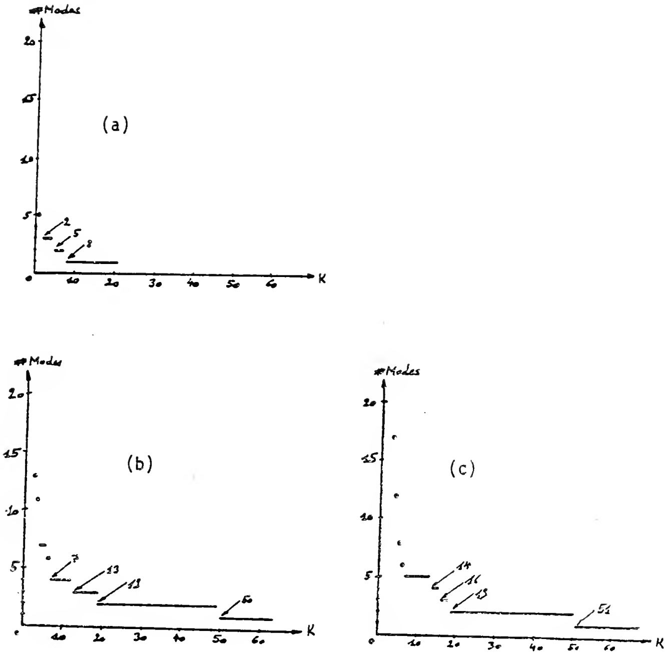

FIGURE F

Diagnostic Plots

(a) chondrite data (1)
(b) iris data, 2 species (2)
(c) iris data, 3 species (3)

TABLE 1(a)

Sample /	100 observations from N(0,1)		50 observations from N(0,1) and 50 fram N(4,1)			
	Ho: 1 mode		Ho: 1 mode		Ho: 2 mode	
# /Distribution	ko	P-value(%)	ko	P-Value(%)		
1	13	97.5	51	0.0	19	14.2
2	18	76.7	45	0.0	18	19.2
3	24	27.5	49	0.0	10	66.7
4	19	56.7	49	0.0	12	63.3
5	19	62.5	46	1.7	16	45.8
6	15	81.7	48	0.8	14	30.8
7	13	95.8	49	2.5	17	33.3
8	17	81.7	45	6.7	14	42.5
9	19	79.2	49	0.0	15	45.8
10	18	65.0	43	5.8	19	50.0
11	17	73.3	49	0.0	17	33.3
12	20	70.0	49	0.0	13	40.8
13	15	95.0	52	0.0	17	21.7
14	18	76.7	48	2.5	21	3.3
15	14	95.0	48	0.0	11	75.0
16	18	58.3	45	2.5	12	62.5
17	22	77.5	50	0.8	12	61.7
18	15	95.0	46	1.7	10	83.3
19	28	35.0	48	3.3	8	98.3
20	19	53.3	45	0.8	12	83.3
21	17	68.3	48	0.0	17	28.3
22	17	72.5	45	1.7	13	70.8
23	19	86.7	48	0.0	13	42.5
24	29	31.7	45	3.3	17	19.2
25	18	64.2	47	1.7	15	60.0

TABLE 1(b)

Sample/	25 observations from N(0,1),		25 from N(4,1)		and 50 from N(8,1)	
	Ho: 1 mode		Ho: 2 mode		Ho: 3 mode	
# /Distribution	ko	P-value(%)	ko	P-value(%)	ko	P-value (%)
1	23	35.0	18	36.7	11	60.8
2	25	4.2	19	40.0	10	72.5
3	23	37.5	22	8.3	20	0.8
4	24	32.5	22	17.5	9	65.0
5	26	3.3	20	20.8	8	93.3
6	23	20.0	19	38.3	15	29.2
7	24	14.2	19	29.2	14	33.3
8	21	35.8	20	13.3	10	75.0
9	26	3.3	25	0.0	11	63.3
10	23	13.3	19	26.7	11	65.0
11	24	25.8	22	7.5	12	77.5
12	23	31.7	22	4.2	9	75.8
13	26	3.3	25	0.0	8	93.3
14	26	15.8	24	0.0	10	78.3
15	26	12.5	23	12.5	13	17.5
16	23	21.7	19	8.3	9	96.7
17	24	55.8	23	7.5	16	19.2
18	20	55.8	16	58.3	9	98.3
19	23	9.2	22	0.8	10	61.7
20	24	20.8	17	45.8	11	53.3
21	24	24.2	16	54.2	10	53.3
22	24	18.3	23	4.2	10	66.7
23	24	20.8	23	4.2	10	62.7
24	23	28.3	20	30.0	13	56.7
25	25	10.8	23	4.2	21	0.8

TABLE 1(c)

Sample/	25 observations from N(0,1), 50 from N(4,1) and 25 from N(8,1)					
	Ho: 1 mode		Ho: 2 modes		Ho: 3 modes	
# / Distribution	ko	P-value (%)	ko	P-value(%)	ko	P-value(%)
1	24	82.5	22	9.2	12	39.2
2	25	62.5	23	4.2	10	71.7
3	25	88.3	24	1.7	20	0.8
4	37	7.5	20	30.8	10	52.5
5	27	77.5	24	1.7	8	97.5
6	24	55.0	23	7.5	15	29.2
7	25	60.2	23	4.2	14	31.7
8	39	6.7	20	29.2	10	71.7
9	26	66.7	23	10.0	11	57.5
10	24	86.7	21	13.3	11	68.3
11	33	19.2	19	25.0	12	65.0
12	26	78.3	23	3.3	9	70.0
13	27	70.0	25	1.7	8	90.0
14	27	51.7	25	2.5	10	75.8
15	31	32.5	22	11.7	13	35.8
16	29	25.8	16	56.7	15	21.7
17	30	53.3	19	50.8	16	15.8
18	23	85.0	21	12.5	9	90.8
19	24	84.2	23	4.2	10	61.7
20	33	35.0	19	30.0	9	79.2
21	23	80.0	21	5.0	10	52.5
22	27	53.3	22	17.5	10	61.7
23	30	50.0	12	97.5	11	39.2
24	25	54.2	24	3.3	13	32.5
25	26	67.7	23	10.0	21	0.0

REFERENCES

Anderberg, M.R. (1973). Cluster Analysis for Applications. New York: Academic Press.

Blashfield, R.K., and Aldenderfer, M.S. (1978). "The Literature on Cluster Analysis". Multivariate Behavioral Research, 13, 271-295.

Cormack, R.M. (1971). "A Review of Classification". Journal of the Royal Statistical Society, Series A, 134, 321-367.

Efron, B. (1979). "Bootstrap methods - another look at the jack-knife." Annals of Statistics, 7, 1-26.

Engelman, L., and Hartigan, J.A. (1969). "Percentage Points of a Test for Clusters". Journal of the American Statistican Association, 64, 1647-1648.

Everitt, B.S. (1974). Cluster Analysis, Halsted Press, New York: John Wiley.

Fisher, R.A. (1936). "Use of multiple measurements in taxonomic problems." Ann Egen, London, 7, 179-188.

Giacomelli, F., Wiener, J., Kruskal, J.B., Pomeran, J.W., and Loud, A.V. (1971). Subpopulations of blook lymphocytes demonstrated by quantitative cytochemistry. Journal of Histochemistry and Cytochemistry 19, 426-433.

Good, I.J. and Gaskin, R.A. (1980). "Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data." Journal of American Stat. Assoc. 75, 42-73.

Hartigan, J.A. (1975). Clustering Algorithms. New York: John Wiley & Sons.

__________(1977). "Clusters as modes", in First International Symposium on Data Analysis and Informatics, Vol. 2, IRIA, Versailles.

__________(1978). "Asymptotic distributions for clustering criteria." Annals of Statistics, 6, 117-131.

__________(1981). "Consistency of single linkage for high-density clusters". Journal of the American Statistical Association, 76, 388-394.

Lee, K.L. (1979). "Multivariate Tests for Clusters." Journal of the American Statistican Association, 74, 708-714.

Silverman, B. (1981). "Using kernel density estimates to investigate multimodality". Journal of the Royal Statistic Society, Series B, 43, 97-99.

Sneath, P.H.A., and Sokal, R.R. (1973). Numerical Taxonomy. San Francisco W.H. Freeman.

Sorenson, T. (1948). "A method of estimating groups of equal amplitude in plant sociology based on similarity of species content". K. Dansek Vidensk, Selsk. Skr. (Biol.), 5, 1-34.

Spath, H. (1980). Cluster Analysis Algorithms. Halsted Press, New York: John Wiley.

Wong, M.A., and Lane (1981). "A kth nearest neighbour clustering procedure". Proceedings of the 13th Interface Sympsium on Statistics and Computer Science, W.F. Eddy (Editor), Springer-Verlag, 308-311.